N88-17243

THE EMPIRICAL ACCURACY OF UNCERTAIN INFERENCE MODELS

David S. Vaughan
Robert M. Yadrick
Bruce M. Perrin
McDonnell Douglas Astronautics Company
P. O. Box 516
St. Louis, MO 63166

Ben P. Wise
Thayer School of Engineering
Dartmouth College
Hanover, NH 03755

ABSTRACT

Uncertainty is a pervasive feature of the domains in which expert systems are designed to function. Several methods have been used for handling uncertainty in expert systems, including probability-based methods, heuristics such as those implemented in MYCIN, methods based on fuzzy set theory and Dempster Shafer theory, and various other schemes. This paper reviews research designed to test uncertain inference methods for accuracy and robustness, in accordance with standard engineering practice. We have conducted several studies to assess how well various methods perform on problems constructed so that correct answers are known, and to find out what underlying features of a problem cause strong or weak performance. For each method studied, we have identified situations in which performance is very good, but also situations in which performance deteriorates dramatically. Over a broad range of problems, some well-known methods do only about as well as a simple linear regression model, and often much worse than a simple independence probability model. Our results indicate that some commercially available expert system shells should be used with caution, because the uncertain inference models that they implement can yield rather inaccurate results.

INTRODUCTION

Uncertainty is a pervasive feature of many domains in which artificially intelligent expert systems are intended to function. Researchers in artificial intelligence have proposed a variety of approaches to uncertain reasoning. Some (e.g., 1, 2, 3) have developed methods that are explicitly based on probability theory. Other approaches, such as those used in MYCIN (4, 5), PROSPECTOR (6), and AL/X (7), use heuristics designed to approximate probability theory. Yet other methods involve adaptations of fuzzy set theory (8), Dempster-Shafer theory (9), and other ideas not based on probability. Unfortunately, there is no wide consensus concerning which approach is best or even suitable for any particular application.

Some researchers have attempted to compare these various approaches through theoretical analysis. For example, Heckerman (10) has shown that the equations which define MYCIN's certainty factors can be translated into probabilistic terms. As another example, Hunter (11) has investigated conditions under which probability theory and Dempster-Shafer theory agree.

Although theoretical analyses can provide useful insights, they also become exceedingly complex and their usefulness for the average practitioner can decrease, particularly when heuristics which have no particular theoretical justification are being considered. Furthermore, these analyses typically focus on the formal assumptions of the various uncertainty models, which are seldom met in practice. Perhaps the most important questions concern how models behave when assumptions are not met.

The present authors and various additional coauthors have taken a different, empirical approach to examining the accuracy of uncertain inference models in a series of studies. We started working independently, but eventually realized the commonalities in our work and began to collaborate.

It should be made clear that we are examining the basic inference models used by systems such as MYCIN. We are not evaluating any particular implementation that uses any given model. In our general approach, answers provided by probability theory are used as a norm against which the accuracy of other uncertain inference models may be measured. These studies differ in details, but all use the same basic research paradigm. First, example inference networks are constructed so that all relevant parameters are known. Next, new values are assigned to the evidence nodes, as though additional information in the form of updated estimates is being supplied by a user during a consultation session. Conclusion node certainty values are calculated which reflect the new information according to the model under consideration. Finally, these answers are compared to results obtained from a probability-based method which provides the minimum cross-entropy solution (12). This approach parallels methods used in various scientific and engineering disciplines, such as sensitivity analysis and "Monte Carlo" simulations, for investigating the behavior of complex systems when assumptions are violated.

We have completed studies which evaluate the MYCIN model, the PROSPECTOR model, probability-based models which contain simplifying assumptions (e.g., independence) and a simple linear model. The objectives of the present paper are to review and summarize these studies, describe the major objectives and findings of each, and discuss the overall implications of these findings for expert system construction and future research.

Table I summarizes the studies that will be reviewed. All of these studies used the general method described above. However, they differed in certain important respects as well. Some focused on only a single uncertain inference model, while others looked at several models simultaneously. Some used many small, randomly-created inference nets, while others used larger, selected nets. Finally, some of these studies derived model parameters by using published theoretical definitions to translate the nets directly, while others "tuned" the parameter values. These issues are all explained and discussed in more detail below.

STUDIES USING THEORETICAL PARAMETERS

In our methodoloy, inference nets are created, solved by a minimum cross-entropy extension of probability theory, and also solved by another uncertain inference model. A key part of this process involved translating between parameters suitable for the probability calculations and parameters required by the other model. For example, the MYCIN model expresses rule strengths (relationships between evidence and conclusions) in measures of believe (MBs) and measures of disbelief (MDs). The developers of MYCIN provided theoretical definitions of these parameters in probability terms. In the first three studies shown in Table I, such theoretical definitions were used for the necessary translations. Consider these studies:

STUDY	UNCERTAIN INFERENCE MODELS	INFERENCE NETS	MODEL PARAMETER ESTIMATION
Wise (13)* Wise & Henrion(14) Wise (15)	MYCIN probability with assumptions	large, selected	theoretical definition
Perrin, Vaughan, Yadrick, Holden, & Kempf (16)	MYCIN	many, small	theoretical definition
Yadrick, Perrin, Vaughan, Holden, & Kempf (17)	PROSPECTOR	many, small	theoretical definition
Wise, Perrin, Vaughan & Yadrick (18)	MYCIN PROSPECTOR probability with assumptions linear regression	many, small	tuned
Wise (19)	PROSPECTOR probability with assumptions linear regression	many, small, selected	tuned

* Wise & Henrion (14) and Wise (15) both contain summaries of results which are presented in more detail in Wise (13).

Table I Evaluation Studies Reviewed

o Wise

Wise (13) presented a detailed theoretical analysis of the MYCIN model, as well as several other models. He also discussed in detail the rationale for accepting the minimum cross-entropy probability solution as an appropriate criterion for evaluating other uncertain inference models. Highlights of this work appear in Wise & Henrion (14), which presents the methodology and some preliminary results, and in Wise (15), which summarizes results for the MYCIN model.

The MYCIN model (5) was one of the first to be used for handling uncertainty in an expert system. It was designed to solve some problems that the developers believed made probability theory unsuitable for their application. The model was also designed to approximate probability calculations while being modular, computationally efficient, and more natural for their subject matter experts to use. The original concerns the developers had with probability theory are probably not valid (e.g., 20). However, the model and several variants are widely cited and used today, particularly in several commercial "shells". Thus, information about the accuracy of the MYCIN model continues to have practical relevence for a large community.

Based on his detailed theoretical analyses and on critical examples cited in the literature, Wise constructed some sets of inference nets with associated rule strengths (defined as probabilities) for which the MYCIN model was predicted to be reasonably accurate, and some for which large errors were predicted. The resulting 30 nets ranged in size from the simplest (two evidence nodes and one conclusion node) to nets with three evidence nodes, multiple conclusion nodes and an intermediate node level. One net comprised nine evidence nodes, four intermediate nodes, and four conclusion nodes. Correlations between pieces of evidence were also varied systematically between strong positive and strong negative associations. With two exceptions, rules in the nets were conjunctive ("AND") rules. To generate test problems, he systematically varied "updated" or input evidence probabilities over four values for each evidence node. This means, for example, that a net with three evidence nodes yielded 64 (4x4x4) problems. Each problem was solved using the probability model, the MYCIN model, and several models based on probability theory with simplfying assumptions. For each inference net he computed $\underline{e}$, the mean squared difference between the maximum entropy probability answers and the inference model answers across the set of problems.

The MYCIN model was most accurate for cases in which there was very little difference between the base rate (prior probability) of the conclusion and the conditional probability of the conclusion when both pieces of evidence were false (or absent). For example, the value of $\underline{e}$ .0004 in one such case and .0005 in another. Conversely, MYCIN was most inaccurate when there was a large difference between the conclusion base rate and the conditional probability of the conclusion when both pieces of evidence were false. The value of $\underline{e}$ was .03, .09, .03, and .04 in four such cases. These results are attributable to two features of the MYCIN model. First, based on theoretical definitions, MYCIN ignores negative evidence. That is, if the updated probability for a piece of evidence is greater than the prior probability (base rate) for that evidence, MYCIN updates conclusion probabilities associated with the evidence. However, if the updated probability for a piece of evidence is below the base rate, MYCIN ignores that information and conclusion probabilities are not updated. The other feature concerns the method used to combine evidence for conjunctive (AND) rules. This method "pays attention to" only one of the pieces of evidence involved. As a consequence of these features, MYCIN provides accurate answers when the impact of ignoring negative evidence is minimized, i.e., when the conditional probability of the conclusion is high given that the evidence is absent.

The robustness of an uncertain inference model can be assessed by examining reasons for its worst performance. In this light, Wise compared the MYCIN model to a simple probability model which assumes conditional independence. Across the set of nets he studied, the conditional probability model was considerably more robust (largest $\underline{e}$ = .04) than the MYCIN model (largest $\underline{e}$ = .09). The MYCIN model was very accurate on some nets, but very inaccurate on other nets.

o Perrin, et al.

Perrin, Vaughan, Yadrick, Holden, and Kempf (16) also studied the MYCIN model. However, the inference nets were constructed in a somewhat different way. In this study, only the simplest sorts of nets were studied, i.e., those comprising two pieces of evidence and one conclusion. These networks are the basic building blocks of larger networks; inference in these nets requires both evidence combining and propagation. In this study, many nets were constructed by random sampling from the universe of three-node nets. In particular, 200 nets were compiled in which the pieces of evidence were independent and 200 were compiled in which the pieces of evidence were statistically associated. Problems were generated by independently varying the updated evidence probabilities over five values. Since all nets had two evidence nodes, this created 25 problems for each net. As in (13), each problem was also solved using the minimum cross-entropy probability model, Next, each net was translated into MYCIN parameters using the theoretical definitions, and was solved using all three of MYCIN's combining functions (conjuctive, disjunctive, and incremental).

Networks were classified according to which combining function provided the lowest error for the network. The incremental function was the most accurate for about 60% of the networks, the conjuctive function was the most accurate for about 35%, and the disjuctive function was most accurate for the remaining 5%. The mean absolute error across all nets was about .07, while the average maximum error per net was about .22. Further analysis indicated that much of this error was due to MYCIN's ignoring negative evidence. For only problems in which MYCIN updated conclusion probabilities, the mean error across problems and nets was about .02, and the average maximum error was about .05. Further analysis indicated that MYCIN error was greatest in these problems when evidence base rates were low and evidence-conclusion associations were strong. These attributes characterize the difficult diagnostic process; the results suggest that MYCIN will be least accurate in precisely the situations for which expert systems are likely to be most valuable.

o Yadrick, et al.

Yadrick, Perrin, Vaughan, Holden, & Kempf (17) studied the model used in the PROSPECTOR system (6). Like the MYCIN model, the PROSPECTOR model was developed to address perceived problems with probability theory for expert system applications. It was also intended to approximate probability calculations while being computationally efficient and modular. While this model has received less attention than the MYCIN model, it and several variants (e.g., AL/X) have also been implemented in commercially-available expert system shells.

Yadrick, et al. used the same inference net and problem generation methods as Perrin, et al. All networks contained two evidence nodes and one conclusion node. A total of 400 networks were sampled which contained independent evidence and 400 nets were sampled which contained associated evidence. Again, 25 problems were generated for each net and solved using maximum entropy probability calculations. The problems were translated into PROSPECTOR parameters using theoretical definitions and the problems were solved using PROSPECTOR conjunctive, disjunctive, and independent rule combining functions. For each net, the mean squared error was computed and the maximum error for a single problem was recorded.

PROSPECTOR error was quite large (often greater than .5) for many nets. Extremely large errors were found mainly for nets in which the probability of the conclusion was high if one piece of evidence was true and one was false, but was not as high if both pieces of evidence were either true or false We concluded that the PROSPECTOR model is fundamentally incapable of handling these "counterintuitive" nets, and excluded them from futher analysis. This left 66 independent and 73 associated evidence nets for additional consideration. The independence combining function was most accurate for about 90% of the remaining independent nets and about 80% of the remaining associated nets. The overall average error was about .014 for independent nets and about .022 for associated nets; overall maximum error was about .055 for independent nets and about .083 for associated nets. Further analysis indicated that error was greatest when the evidence is most strongly associate with the conclusion. Moreover, the error can be compounded or mitigated by the values of updated evidence probabilities. In summary, the PROSPECTOR model was quite accurate for some problems and networks, but very inaccurate for others over a wide range of new evidence probabilities. Like MYCIN, it appears to be least accurate in the typical situations to which it would likely be applied.

TUNED MODEL PARAMETERS

The studies described above used published formal definitions to translate between probability model parameters and uncertainty model parameters. The hope was to determine the absolute degree of error and provide a theoretical explanation for sources of error produced by the uncertainty models. Despite some success in this, practical applications of the findings are limited, precisely because we used formal uncertainty model parameter definitions. These theoretical definitions have little relevence to knowledge engineers building real expert systems, because parameters are typically estimated by experts based on an intuitive rather than a formal understanding. Then the parameters are "tuned", or adjusted interactively by the experts and knowledge engineers to obtain the most accurate results on the data used for system development. The relationship between parameters estimated in this way, the formal definitions of the parameters, and probability theory is not clear. Furthermore, the tuning process may correct some or all of the errors observed in the studies described above.

This tuning issue lead us to do two additional studies (18, 19). The objective was to study the errors made by uncertain inference models empirically after their parameters have been tuned. As before sample networks were created, and problems were run by systematically varying updated evidence probabilities. Problem solutions produced by uncertain inference models were compared to the same minimum cross-entropy probability norm. This time, however, the model parameters were optimized for each net ("tuned") so that the model's answers were as close to the probability answers as possible, on the average. These solutions, therefore, represent the best performance that could be achieved by each model.

o Wise, et al.

Four different inference methods were examined by Wise, Perrin, Vaughan, & Yadrick (18).. These included the MYCIN and PROSPECTOR models as before. We also included a linear regression model, given by the following equation:

$$P'(C) = a + b1*P'(E1) + b2*P'(E2). \tag{1}$$

In this equation and below, P'(x) is the updated probability for the event x, and a, b1, and b2 are constant parameters, which were optimized. Linear models have received little attention from artificial intelligence researchers, although they have been used successfully to model a variety of human judgments (21). We included this model to provide a baseline against which to compare the other models.

Finally, a probability theory-based independence model was also included. This model is described by the following equation:

$$\begin{aligned} P'(c) = {} & P'(\sim E1) * P'(\sim E2) * P(C|\sim E1\&\sim E2) + \\ & P'(E1) * P'(\sim E2) * P(C|E1\&\sim E2) + \\ & P'(\sim E1) * P'(E2) * P(C|\sim E1\&E2) + \\ & P'(E1) * P'(E2) * P(C|E1\&E2). \end{aligned} \tag{2}$$

The model reflects normal probability calculations under the assumption that the pieces of evidence are independent. The four conditional probabilities are the model parameters (which were optimized). After parameter optimization, this model is equivalent to a linear regresion model with an interaction term.

A total of 109 two-evidence, one-conclusion networks were sampled using procedures similar to those of (16) and (17). For each piece of evidence, the updated probabilities varied over five values so that 25 problems were run for each network. For each model, parameter values were obtained which minimized, across the 25 problems, the sum of squared differences between the model solutions and the minimum cross-entropy answers. This optimization was done using a deflected gradient search algorithm (22) with appropriate precautions to avoid local minima and round-off error problems. Table II summarizes the performance of the four models.

INFERENCE METHOD	AVERAGE RMSE	HIGH RMSE	LOW RMSE
MYCIN	.048	.152	.001
PROSPECTOR	.047	.148	.001
Linear eq.	.048	.152	.001
Independence	.006	.036	.000

Note: This table was taken from (18). RMSE is root mean squared error.

Table II Tuned Parameter Errors

The main finding was that the MYCIN (5 parameters), PROSPECTOR (7 parameters), and linear (3 parameters) models performed equally well (for all practical purposes), while the independence (4 parameters) model was significantly more accurate (according to an analysis of variance test). Furthermore, the errors for the MYCIN, PROSPECTOR, and linear models were highly correlated (Pearson product-moment coefficient >.95). This shows that the models all performed well or poorly on the same problems. They were behaving almost identically for the networks studied here, although the linear model requires estimation of fewer parameters. A probability theory-based independence model performed better and required fewer parameters than MYCIN or PROSPECTOR.

o Wise

The objective of this study (19) was to determine the degree to which errors of the sort shown in Table I can be attributed to assumption violations in the networks. The study included the PROSPECTOR model, the linear and marginal independence models (equations 1 and 2), and a model that was linear on logarithms of odds ratios (i.e., it substituted logs of odds ratios for probabilities in equation 1). The general methodology, including network and problem generation and model parameter optimization, were the same as (18). Here, however, all networks were constructed to meet the PROSPECTOR model's conditional independence assumptions. Thus, all error for PROSPECTOR in these networks must be due to the approximate updating functions.

Table III summarizes results for the conditional independence networks. In this table, errors are expressed in terms of a standardized measure, where 1.0 reflects no error and 0.0 reflects the same level of error as the linear model. Positive scores on this measure indicate better performance than the linear model, and negative scores indicate worse performance than the linear model. As

	INDEPENDENCE	PROSPECTOR	LOG-ODDS
Mean	.90	.52	-.36
Standard-ized error	.08	.35	.07

Note: This table summarized from (X). Cell entries are standardized error measure (see text).

Table III Errors for Conditionally Independent Networks

may be seen, the PROSPECTOR model performs better than the linear model when networks meet the model's assumptions. However, due to the updating procedure, the model still performs worse than the independence model and still makes substantial errors.

DISCUSSION

We think that these results have a number of implications for expert system construction.

First, it is clear that both the MYCIN and PROSPECTOR models are suitably accurate under some circumstances, but can make large errors under other circumstances. Exactly which model and what parameter values are used make a potentially important difference in the overall accuracy of an expert system. It may not be possible to tune the system to perform with reliable accuracy across a broad range of problems, users, and solutions. In short, under some circumstances one should probably not use the MYCIN or PROSPECTOR models. This conclusion is important, since these models are embedded in many commercial shells and are widely used. Indeed, neither these nor other models should be used uncritically, without investigations to determine appropriateness to the particular application under consideration.

Second, very simple models may work well for many problems. A simple linear model worked as well as the MYCIN and PROSPECTOR models, and a probability-based independence model worked much better. Elaborate models have been developed to handle uncertainty in expert systems, but the elaborations add little to accuracy and are very sensitive to differences in, for example, evidence-conclusion relationships.

All of the uncertain inference models made substantial errors under some circumstances. This suggests that for some difficult applications, custom-built uncertain inference models may still be required. The system builder should select or develop a method that is neither too simple nor too complex for the application at hand.

When an uncertain inference model is being considered, one need not focus entirely on the assumptions of the model and whether those assumptions are met in the application. We have found that some models work well even when assumptions are not met (e.g., the probability-based independence model and the linear model) and that others may work poorly even if assumptions are met. We believe that robustness is more important than theoretical elegance in practical expert system building.

Finally, we believe that the empirical approach to evaluating uncertain inference model accuracy and the general methodology we have developed is useful. The findings summarized above have shed new light on the performance of such models, which goes beyond theoretical analyses. However, many questions remain unanswered. Our studies have looked at only a few models and only at simple networks. While it seems likely to us that errors will tend to propagate and compound in many large networks, that other heuristic models will perform poorly in many circumstances, these issues should be settled empirically. We are presently investigating these and other issues.

Acknowledgment — Some of the research reviewed in this paper was conducted under the McDonnell Douglas Independent Research and Development program.

REFERENCES

1. Pearl, J., "How to do with probabilities what people say you can't", Proceedings of the IEEE Second Conference on Artificial Intelligence Applications, Miami, FL, 1985.

2. Cheeseman, P., "A method of computing generalized Bayesian probability values for expert systems", Proceedings of the Eighth International Joint Conference on Artificial Intelligence, Karlsruhe,
West Germany, 1983.

3. Vaughan, D.S., Perrin, B.M., Yadrick, R.M., Holden, P.D., and Kempf, K.G. "An odds ratio based inference engine," UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, Kanal, L.N. & Lemmer, J.F. (Eds.), North-Holland, Amsterdam, 1986. 383-389.

4. Shortliffe, E.H. & Buchanan, B.G., "A model of inexact reasoning in medicine," MATHEMATICAL BIOSCIENCES, 23, 351-379, 1975.

5. Shortliffe, E.H., COMPUTER-BASED MEDICAL CONSULTATION: MYCIN, American Elsevier, Amsterdam, 1976.

6. Duda, R.O., Hart, P.E., Konolige, K., & Reboh, R., "A computer-based consultant for mineral exploration, Final Report, Project 6415, SRI International, Menlo Park, CA, 1979.

7. Reiter, J., "AL/X: An expert system using plausible inference," Intelligent Terminals, Ltd., Oxford, 1980.

[8] Zadeh, L.A., "The role of fuzzy logic in the management of uncertainty in expert systems," MEMORANDUM # UCB/ERLM83/41, Electronics Research Laboratory, University of California, Berkeley, CA, 1983.

9. Shafer, G., "Probability judgment in artificial intelligence", UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, Kanal, L.N. & Lemmer, J.F. (Eds.), North-Holland, Amsterdam, 1986, 127-135.

10. Heckerman, D. "Probabilistic interpretation for MYCIN's certainty factors," UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, Kanal, L.N. & Lemmer, J.F. (Eds.), North-Holland, Amsterdam, 1986, 167-196.

11. Hunter, D., "Dempster-Shafer vs. Probabilistic Logic", Proceedings of the Third AAAI/Martin Marietta/ADS Workshop on Uncertainty in Artificial Intelligence, Seattle, WA, July, 1987.

12. Shore, J.E. & Johnson, R.W., "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," IEEE TRANSACTIONS ON INFORMATION THEORY, Vol. IT-26, 1980, 26-37.

13. Wise, B.P., "Experimental comparison of uncertain inference systems," Unpublished Ph.D. Dissertation, Department of Engineering and Public Policy, Carnegie-Mellon University, Pittsburgh, PA, 1986.

14. Wise, B.P. & Henrion, M., "A framework for comparing uncertain inference systems to probability," UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, Kanal, L.N. & Lemmer, J.F. (Eds.), North-Holland, Amsterdam, 1986, 69-83.

15. Wise, B.P., "Experimentally comparing uncertain inference systems to probability," Proceedings of the Second RCA/AAAI Workshop on Uncertainty in Artificial Intelligence, Philadelphia, PA, August, 1986. To appear in a forthcoming book from North-Holland.

16. Perrin, B.M., Vaughan, D.S., Yadrick, R.M., Holden, P.D., & Kempf, K.G., "Evaluation of uncertain inference models II: MYCIN, MDC TECHNICAL REPORT # E3205, McDonnell Douglas Corporation, St. Louis, MO, July, 1987.

17. Yadrick, R.M., Perrin, B.M., Vaughan, D.S., Holden, P.D., & Kempf, K.G., "Evaluation of uncertain inference models I: PROSPECTOR," Proceedings of the Second RCA/AAAI Workshop on Uncertainty in Artificial Intelligence, Philadelphia, PA, August, 1986. To appear in a forthcoming book from North-Holland.

18. Wise, B.P., Perrin, B.M., Vaughan, D.S., & Yadrick, R.M., "The role of tuning uncertain inference systems," Proceedings of the Third AAAI/Martin Marietta/ADS Workshop on Uncertainty in Artificial Intelligence, Seattle, WA, July, 1987.

19. Wise, B.P., "Satisfaction of assumptions is a weak predictor of performance," Proceedings of the Third AAAI/Martin Marietta/ADS Workshop on Uncertainty if Artificial Intelligence, Seattle, WA, July, 1987.

20. Cheeseman, P., "In defense of probability," Proceedings of the Ninth International Joint Conference on Artificial Intelligence, Los Angeles, CA, August, 1985.

21. Dawes, R.M., "The robust beauty of improper linear models in decision making," AMERICAN PSYCHOLOGIST, 34, 571-582.

22. Beightler, C.S., Phillips, D.T., & Wilde, D.J., "FOUNDATIONS OF OPTIMIZATION, Prentice-Hall, Englewood Cliffs, NJ, 1979.